Quantile-based Distributional RL is a Special Case of Quantile MDP with Markovian Policy

Yudong Luo

May 29, 2026

1 Notation and General Assumptions

We denote the random variable by a tilde, i.e., $\tilde{x}$, denote the augmented real as $\bar{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$.

Quantiles and Value-at-Risk. The quantile of random varible $\tilde{x}$ at level $\alpha \in [0, 1]$ is any $\tau \in \mathbb{R}$ such that $\mathbb{P}[\tilde{x} \leq \tau] \geq \alpha$ and $\mathbb{P}[\tilde{x} < \tau] \geq 1 - \alpha$. It might not be unique and lies in the interval $[q_\alpha^-(\tilde{x}), q_\alpha^+(\tilde{x})]$, where

$$q_\alpha^-(\tilde{x}) := \min\{\tau \in \bar{\mathbb{R}} | \mathbb{P}[\tilde{x} \leq \tau] \geq \alpha\}$$
$$q_\alpha^+(\tilde{x}) := \max\{\tau \in \bar{\mathbb{R}} | \mathbb{P}[\tilde{x} < \tau] \leq \alpha\}.$$

For ease of presentation, we assume the quantile is unique, i.e., $q_\alpha^-(\tilde{x}) = q_\alpha^+(\tilde{x})$, for any $\tilde{x}$. In this case, value-at-risk (VaR) equals quantile.

2 Quantile-based Distributional RL

Distributional reinforcement learning is getting popular since the seminal work of [Bellemare et al., 2017], where a distributional Bellman equation is proposed, i.e.,

$$\tilde{z}(s, a) \overset{d}{=} \tilde{r}(s, a) + \gamma \tilde{z}(\tilde{s}', \tilde{a}') \tag{1}$$

In this equation, $\tilde{z}(s, a) := \sum_{t=0}^{\infty} [\gamma^t \tilde{r}(\tilde{s}_t, \tilde{a}_t) | \tilde{s}_0 = s, \tilde{a}_t = a]$ is the random variable of the return, and $\overset{d}{=}$ means equal in distribution. Note that Eq. 1 is more like policy evaluation where $\tilde{a}'$ comes from $\pi(\cdot|\tilde{s}')$. [Bellemare et al., 2017] mentioned that in control setting, the distributions on the two hand side of Eq. 1 might not be equal.

[Bellemare et al., 2017] used categorical distribution to represent the value distribution. Later, [Dabney et al., 2018] proposed to represent the value distribution by its empirical inverse CDF (quantile function), and update the quantile estimates by quantile regression given sampled (s, a, r, s') as

$$q_\alpha(s, a) \leftarrow q_\alpha(s, a) - \eta \cdot \partial_y \mathbb{E}_{u \sim U[0,1]} \Big[l_\alpha\big(r + q_u(s', a') - y\big) \Big] \Big|_{y = q_\alpha(s,a)}, \ \forall \alpha \in [0, 1] \tag{2}$$

where $q_\alpha(s, a) := \mathrm{VaR}_\alpha[\tilde{z}(s, a)]$ represents the α quantile of the state-action value; η is the learning rate; $U[0, 1]$ is a uniform distribution on $[0, 1]$; $l_\alpha(\cdot)$ is the loss function corresponds to quantile regression given by

$$l_\alpha(x - y) := (\alpha - \mathbb{I}\{x < y\})(x - y);$$

a' can come from $\pi(\cdot|s')$ if doing policy evaluation, and $a' = \arg\max_a \mathbb{E}_u[q_u(s', a)]$ when performing optimal control.

Quantile-based distributional RL approach has become one of the main stream approaches since it was introduced. Recently, [Rowland et al., 2024] provided theoretical analysis for quantile TD learning (Eq. 2).

3 Quantile MDP

Consider a general problem, given a Markov Decision Process (MDP), what is the optimal quantile value for all $\alpha \in [0,1]$ and (s,a), i.e.,

$$q^*_\alpha(s,a) := \max_{\pi_\mathrm{h}} \mathrm{VaR}^{\pi_\mathrm{h}}_\alpha \Big[\sum_{t=0}^{\infty} \gamma^t \tilde{r}(\tilde{s}_t, \tilde{a}_t) | \tilde{s}_0 = s, \tilde{a}_0 = a \Big]. \tag{3}$$

The corresponding optimal policy of Eq. 3 is history-dependent since VaR_α operator is non-linear, therefore denoted as π_h in Eq. 3.

[Li et al., 2022] and [Hau et al., 2023] showed that this static VaR problem enjoys a dynamic decomposition which admits a Bellman-like equation. However, the equations in [Li et al., 2022] and [Hau et al., 2023] are model-based where a constrained optimization problem involving the transition probability need to be solved. To address this limitation, [Hau et al., 2025] proposed a nested VaR Bellman equation as

$$q^*(s,\alpha,a) = \mathrm{VaR}_\alpha \Big[\tilde{r}(s,a) + \gamma \max_{a'} q^*(\tilde{s}', \tilde{u}, a') \Big], \tag{4}$$

where $\tilde{u}$ follows $U[0,1]$ and VaR_α is applied to the joint distribution of $\tilde{s}'$ and $\tilde{u}$. We have moved the parameter α inside the q function to treat it as an augmentation of the state space.

Notice that α-quantile is the argmin of the quantile regression loss, as a result, Eq. 4 can be alternatively expressed as

$$q^*(s,\alpha,a) = \arg\min_y \mathbb{E} \Big[l\big(\tilde{r}(s,a) + \gamma \max_{a'} q^*(\tilde{s}', \tilde{u}, a') - y\big) \Big]. \tag{5}$$

As a result, the quantile value can be updated by gradient descent towards the decreasing direction of the quantile regression loss. Given sampled (s,α,a,r,s'), the quantile value is updated by

$$q(s,\alpha,a) \leftarrow q(s,\alpha,a) - \eta \cdot \partial_y \mathbb{E}_{u\sim U[0,1]} \Big[l_\alpha\big(r + \gamma \max_{a'} q(s',u,a') - y\big) \Big] \Big|_{y=q(s,\alpha,a)}, \tag{6}$$

where η is the learning rate.

Eq. 2 and Eq. 6 share very similar structure. The only difference is that Eq. 6 obtains the optimal action a' for each quantile-level $u \in [0,1]$, while Eq. 2, in control setting, uses the optimal risk-neutral action $a' = \arg\max_a \mathbb{E}_u[q(s',u,a)]$. However, this subtle difference suggests that quantile-based distributional RL approach (in control setting) is **not** learning the optimal quantile values in an MDP. Another explanation is that the optimal quantile policy is history-dependent while distributional RL uses Markovian policy.

4 Quantile Decomposition under Markovian Policy

We show that when policy is Markovian, the quantile decomposition technique (Li et al. [2022]) used in quantile MDP reduces to quantile-based distributional RL approach in policy evaluation.

For ease of presentation, here we assume the Markovian policy is deterministic, i.e., $\pi(s)$ (so that we do not need to decompose the randomness coming from the policy at this point)

$$
\begin{aligned}
q(s,\alpha,\pi(s)) &= \mathrm{VaR}_\alpha\Big[\sum_{t=0}^{\infty}\gamma^t\tilde{r}(\tilde{s},\pi(\tilde{s}))|\tilde{s}_0=s\Big] \\
&= \mathrm{VaR}_\alpha\Big[\tilde{r}(s,\pi(s))+\gamma\sum_{t=1}^{\infty}\gamma^{t-1}\tilde{r}(\tilde{s}_t,\pi(\tilde{s}_t))\Big] \\
&\overset{(a)}{=} \mathrm{VaR}_\alpha\Big[\sum_{(r_i,s_i')}P(r_i,s_i'|s,\pi(s))\Big(r_i+\gamma\big[\sum_{t=1}^{\infty}\gamma^{t-1}\tilde{r}(s_i',\pi(s_i'))|\tilde{s}_1=s_i'\big]\Big)\Big] \\
&\overset{(b)}{=} \mathrm{VaR}_\alpha\Big[\sum_{(r_i,s_i')}P(r_i,s_i'|s,\pi(s))\Big(r_i+\gamma q(s_i',\tilde{u},\pi(s_i'))\Big)\Big] \\
&\overset{(c)}{=} \max_{\boldsymbol{\beta}}\min_i \mathrm{VaR}_{\beta_i}[r_i+\gamma q(s_i',\tilde{u},\pi(s_i'))] \;\text{ with }\; \sum_i \beta_i\cdot P(r_i,s_i'|s,\pi(s))\le\alpha \\
&= \max_{\boldsymbol{\beta}}\min_i r_i+\gamma q(s_i',\beta_i,\pi(s')) \\
&\overset{(d)}{=} \mathrm{VaR}_\alpha\Big[\tilde{r}+\gamma q(\tilde{s}',\tilde{u},\pi(\tilde{s}'))\Big],
\end{aligned}
$$

where (a) considers all the possible combinations of (r, s') under transition $P(r, s'|s, a)$; (b) replaces the distribution of $\big[\sum_{t=1}^{\infty}\gamma^{t-1}\tilde{r}(s_i',\pi(s_i'))|\tilde{s}_1=s_i'\big]$ by its equivalent representation $q(s_i', \tilde{u}, \pi(s_i'))$; (c) is according the quantile decomposition theory, i.e, theorem 1 and Lemma 2 in [Li et al., 2022]; (d) is according to Lemma B.4 of [Hau et al., 2025].

Following the same idea as Eq. 5 and Eq. 6, we have

$$
q(s,\alpha,\pi(s)) = \arg\min_y \mathbb{E}\Big[l_\alpha\big(\tilde{r}+\gamma q(\tilde{s}',\tilde{u},\pi(\tilde{s}'))-y\big)\Big],
$$

which leads to the update rule as Eq. 2 given sampled (s, a, r, s').

5 Remark

Note that some technique details are omitted in this note. Please refer to [Hau et al., 2025] for handling the non-smoothness of quantile regression loss, and the discussion when quantile level chooses the boundary value 0 and 1.

References

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pages 449–458. Pmlr, 2017.

Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Jia Lin Hau, Erick Delage, Mohammad Ghavamzadeh, and Marek Petrik. On dynamic programming decompositions of static risk measures in markov decision processes. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:51734–51757, 2023.

Jia Lin Hau, Erick Delage, Esther Derman, Mohammad Ghavamzadeh, and Marek Petrik. Q-learning for quantile mdps: A decomposition, performance, and convergence analysis. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025.

Xiaocheng Li, Huaiyang Zhong, and Margaret L Brandeau. Quantile markov decision processes. *Operations research*, 70(3):1428–1447, 2022.

Mark Rowland, Rémi Munos, Mohammad Gheshlaghi Azar, Yunhao Tang, Georg Ostrovski, Anna Harutyunyan, Karl Tuyls, Marc G Bellemare, and Will Dabney. An analysis of quantile temporal-difference learning. *Journal of Machine Learning Research*, 25(163):1–47, 2024.